A MAXIMUM LIKELIHOOD MODEL FOR ESTIMATING ORIGIN-DESTINATION MATRICES

by

Heinz Spiess

This research was supported in part by a grant from the National Sciences and Engineering Research Council of Canada.

Université de Montréal
Centre de recherche sur les transports - Publication #293
Département d'informatique et de
recherche opérationnelle - Publication #476
Avril 1983

Dépôt légal - 3e trimestre 1983

ABSTRACT

We describe a model for estimating an origin-destination matrix from an observed sample matrix, when the volumes on a subset of the links of the network and/or the total productions and attractions of the zones are known. The elements of the observed sample matrix are assumed to be integers that are obtained from independent Poisson distributions with unknown means. A maximum likelihood model is formulated to estimate these means, yielding an estimation of the "true" origin-destination matrix which is consistent with the observed link volumes. Conditions for existence and uniqueness of a solution are discussed. A solution algorithm based on the cyclic coordinate descent method is developed and its convergence properties are analyzed. The special case of the matrix estimation problem, in which marginal totals are given instead of link volumes, is considered separately; a numerical example is used to illustrate the problem. Using results about the asymptotic behavior of the distribution of the likelihood function, tests may be derived that allow statistical inference on the consistency of the available data. Finally, an extension of the model is studied in which the observed volumes are Poisson-distributed as well.

RESUME

Nous décrivons un modèle pour l'estimation d'une matrice origine-destination à partir d'une matrice observée quand les volumes sur les liens du réseau et/ou les productions et les attractions totales des zones sont connus. Nous considérons le cas où les élements de la matrice observée proviennent des distributions de Poisson indépendantes avec des moyennes inconnues. Un modèle de vraisemblance maximale est formulé pour trouver une estimation de la "vraie" matrice origine-destination tout en restant consistant avec les volumes observés sur les liens du réseau. Nous étudions les conditions sous lesquelles l'existence et l'unicité d'une solution optimale du modèle sont assurées. Ensuite nous élaborons un algorithme de résolution basé sur la méthode de descente cyclique au long des axes de coordonnées et nous discutons ses propriétés de convergence. Le cas spécial dans lequel les données observées correspondent aux totaux marginaux de la matrice origine-destination est cosidéré séparément et un exemple numérique est utilisé pour illustrer ce problème. En utilisant des résultats sur le comportement asymptotique de la distribution de la fonction de vraisemblance il est possible de construire des tests statistiques qui permettent de faire l'inférence sur la consistance des données utilisées. Finalement on propose une extension du modèle pour traiter le cas dans lequel les volumes observés proviennent eux-mêmes d'une distribution de Poisson.

1. Introduction

The problem of reconstituting an origin-destination matrix from observed link volumes has been studied by several authors, most recently by Jörnsten and Nguyen[4], Van Zuylen and Willumsen[14] and Van Zuylen[12]. Nguyen[9] gives a good overview on the state of the art in this research domain. The general problem can be formulated in the following way. Let P denote the set of origins, Q the set of destinations and I=PxQ the set of origin destination pairs. Most of the existing models to estimate an O-D matrix $\{T_i\}$ may be written in the form

$$\text{Min} \quad f(T,t) \tag{1}$$

subject to

$$\sum_{i \in I} n_{ia} T_i = V_a \qquad ,a \in A, \tag{2}$$

$$T_i \geq 0 \qquad ,i \in I. \tag{3}$$

where the coefficients n_{ia} correspond to the proportion of trips for O-D pair i that use link a, and V_a denotes the observed volume on link a. The matrix $\{t_i, i \in I\}$ contains some a priori information on the O-D matrix $\{T_i\}$. The objective function f(T,t) is either a measure of distance between $\{T_i\}$ and $\{t_i\}$ in some specific metric [2,12], or the widely used convex entropy function [4,9,10,11,14]

$$f(T,t) = \sum_{i \in I} T_i \log \frac{T_i}{t_i}. \tag{4}$$

While these particular forms for f(T,t) are very plausible, the choice of which function is used in a particular case is rather arbitrary. On the one hand, the matrix $\{t_i\}$ has not been given a precise meaning for any of the methods described in the literature. Vague descriptions are used to define $\{t_i\}$ [4,9,11,12], such as "initial estimate", "outdated origin-destination matrix", "a priori guess", "target matrix" and "reference matrix". On the other hand, it has been recognized that "the accuracy of the resulting estimates is highly sensitive to the a priori information contained in the reference input trip matrix $\{t_i\}$" [9]. Given this rather imprecise, not yet well understood and perhaps confusing role of the reference matrix $\{t_i\}$, it is not surprising that the models developed so far all lack means for statistically inferring the validity of the resulting estimates.

In this paper we shall see that it is possible to overcome the above mentioned problems by giving a precise meaning to the matrix $\{t_i\}$. The statistical model chosen restricts the class of problems considered, but it also allows us to gain much more insight into the structure of these problems. In the following, we shall consider the case in which the matrix $\{t_i\}$ is obtained by sampling, for each O-D pair $i \in I$, a Poisson distributed random variable with unknown mean T_i. Thus, $\{t_i\}$ is one observation of a set of random variables $X_i, i \in I$, that are distributed according to independent Poisson distributions with means T_i. Kirby [5]

considers the same sampling distribution for the observed trip matrix when studying the problem of calibrating gravity models. The assumption about the matrix $\{t_i\}$ corresponds to many situations that we find in practice. Most of the existing O-D matrix survey techniques are based on counting individuals and yield therefore observed matrices with integer values. If the survey consists in sampling a small percentage of a finite population with a stable traffic pattern, the observed trip matrix $\{t_i\}$ follows a multinominal distribution. As the sampling percentage becomes small, it is well known that the multinominal distribution may be approximated by independent Poisson distributions. Most "work trip" surveys belong to this category. A Poisson distributed matrix $\{t_i\}$ is also obtained when observing mutually independent events over a finite time interval. Off-peak trips and telephone calls fit well into this framework.

The aim of the model is to estimate the parameters $\{T_i\}$ of the underlying Poisson processes, using the sample matrix $\{t_i\}$, as well as the observed link volumes. Since observed link volumes are normally collected on a day to day basis, it is reasonable to assume that we know their "true" average values or at least very good approximations thereof.

In this paper we only consider the case in which the link proportions n_{ia} are constant and exogenously determined. Without loss of generality we shall further assume that the feasible set defined by (2) and (3) does not res-

trict any variable T_i to a single value. If this were the case, these variables could simply be eliminated. As a consequence, we also assume $V_a>0$, $a\epsilon A$, since $V_a=0$ would force $T_i=0$ for all $i\epsilon I$ with $n_{ia}>0$.

The remaining part of this paper is organized as follows. In Section 2 we formulate the model that is based on the maximum likelihood method and we study some of its mathematical properties. In Section 3 we propose a solution algorithm based on the cyclic coordinate descent method and analyze some of its convergence properties. In Section 4 we examine a special case of the general matrix estimation problem in which instead of link volumes the marginal totals of the O-D matrix (productions and attractions) are known. Section 5 contains some results that allow statistical inference on the validity of the model and in Section 6 the model is extended to the case in which the link volumes are not the true mean values but also Poisson distributed observations. In Section 7 conclusions are drawn.

2. A Maximum Likelihood Model

Under the hypothesis that for each O-D pair t_i is obtained by observing an independent Poisson process with mean T_i, the probability of observing t_i is

$$\text{Prob}[\ \text{Poisson}(T_i)=t_i\ ] = \frac{T_i^{t_i}\, e^{-T_i}}{t_i!} \qquad ,i\epsilon I. \qquad (5)$$

The joint probability of observing the sample matrix (t_i) is therefore

$$\text{Prob}[\ (t_i)\ ] = \prod_{i\epsilon I} \text{Prob}[t_i] = \prod_{i\epsilon I} \frac{T_i^{t_i}\, e^{-T_i}}{t_i!} . \qquad (6)$$

Furthermore we know that since the matrix (T_i) satisfies (2) and (3), it is non-negative and consistent with the observed link flows V_a, $a\epsilon A$.

Applying the maximum likelihood estimation technique [15,16] to this problem amounts to finding the matrix (T_i^*) which satisfies the given constraints and yields the maximum probability (6) for observing (t_i). By replacing the probability (6) with its logarithm and adopting the convention $0 log 0 = 0$, we can formulate the maximum likelihood model as

$$\text{Max} \quad \sum_{i\epsilon I} (t_i log T_i - T_i - log t_i!) \qquad (7)$$

subject to (2) and (3).

Since the last term in (7) is constant, we can drop it

and rewrite (7) as

$$\text{Min} \sum_{i \in I} (T_i - t_i \log T_i) \tag{8}$$

subject to (2) and (3). Hence, we have formulated another problem that belongs to the family (1) with

$$f(T,t) = \sum_{i \in I} (T_i - t_i \log T_i) \tag{9}$$

Note that (8) is convex in T and is strictly convex in T if $t_i > 0$, $i \in I$. If the link volumes are feasible and the set of constraints (2) and (3) is consistent, then the existence of an optimal solution of (8) is assured, since the unconstrained problem has a finite optimal solution $\{T_i\}=\{t_i\}$. The solution to problem (8) may therefore be obtained by any of the standard solution methods for convex programming problems. In section 3 we will develop an algorithm that is especially adapted to this problem.

Let T^* denote an optimal solution of problem (8). As we shall see in the following development, it is important to classify the O-D pairs into two categories, according to whether or not the corresponding element of the observed sample matrix $\{t_i\}$ is empty or not. Thus, let I^+ denote the set of O-D pairs for which $t_i > 0$, and let I^o denote the set of O-D pairs for which $t_i = 0$.

Proposition 1: T_i^*, $i \in I^+$, are unique.

Proof: Problem (8) is convex and separable by O-D pair $i \in I$. Since for $i \in I^+$ the corresponding part of the objective

function is strictly convex, T_i^*, $i \epsilon I^+$, are unique. ■

Proposition 2: Given the optimal values T_i^*, $i \epsilon I^+$, the optimal values T_i^*, $i \epsilon I^o$, may be found by solving the following linear programming problem

$$\text{Min} \sum_{i \epsilon I^o} T_i \tag{10}$$

subject to

$$\sum_{i \epsilon I^o} n_{ia} T_i = V_a - \sum_{i \epsilon I^+} n_{ia} T_i^* \quad , a \epsilon A, \tag{11}$$

$$T_i \geq 0 \quad , i \epsilon I^o. \tag{12}$$

Proof: We first rewrite (8) by separating the O-D pairs according to I^o and I^+

$$\text{Min} \sum_{i \epsilon I^o} T_i + \sum_{i \epsilon I^+} (T_i - t_i log T_i) \tag{13}$$

subject to

$$\sum_{i \epsilon I^o} n_{ia} T_i + \sum_{i \epsilon I^+} n_{ia} T_i = V_a \quad , a \epsilon A \tag{14}$$

and (3). Since we know the optimal values of T_i, $i \epsilon I^+$, we can insert them into (13) and (14). The second term of (13) is now constant and may be dropped, yielding (11). ■

Proposition 3: The total number of trips $\sum_{i \epsilon I} T_i^*$ of an optimal solution of (8) is unique.

Proof: According to Proposition 1 the values of T_i^*, $i \epsilon I^+$, are unique and therefore $\sum_{i \epsilon I^+} T_i^*$ is unique as well. Using Proposition 2, we see that $\sum_{i \epsilon I^0} T_i^*$ is the value of the objective function of problem (11) and is therefore unique. Hence the total number of trips $\sum_{i \epsilon I} T_i^* = \sum_{i \epsilon I^0} T_i^* + \sum_{i \epsilon I^+} T_i^*$ is unique. ∎

The uniqueness of the total number of trips is a significant property of the proposed model, since when $t_i=0$ have been observed for some O-D pairs, the estimated trip matrix T^* may not necessarily be unique. If, on the other hand, the total number of trips is known from other sources, it is possible to include it as an additional constraint in (2), using $n_{ia}=1$, $i \epsilon I$ and V_a equal to the total number of trips.

We now turn our attention to the dual formulation of problem (8). Let L_a, $a \epsilon A$, denote the Lagrangean multipliers that are associated with the constraints (2). The Lagrangean dual of problem (8) is then

$$\max_{L} \{ \inf_{T \geq 0} \{ \sum_{i \epsilon I} (T_i - t_i log T_i) + \sum_{a \epsilon A} L_a (\sum_{i \epsilon I} n_{ia} T_i - V_a) \} \} \qquad (15)$$

By rearranging terms and reversing the order of summation we can rewrite (15) as

$$\max_{L} \{ \inf_{T \geq 0} \{ \sum_{i \epsilon I} T_i (1 + \sum_{a \epsilon A} n_{ia} L_a) - \sum_{i \epsilon I} t_i log T_i - \sum_{a \epsilon A} L_a V_a \} \} \qquad (16)$$

and by noting that the problem of finding the infimum is separable by O-D pair, (16) becomes

$$\max_{L} \Big\{ \sum_{i\in I} \inf_{T_i \geq 0} \{T_i (1 + \sum_{a\in A} n_{ia} L_a) - t_i \log T_i\} - \sum_{a\in A} L_a V_a \Big\}. \qquad (17)$$

The latter formulation makes it easy to discern the conditions that L_a must satisfy in order to be essentially feasible; the infimum in (16) is only finite for multipliers L_a, $a\in A$, that satisfy

$$1 + \sum_{a\in A} n_{ia} L_a \begin{cases} > 0 & \text{, if } t_i > 0 \\ \geq 0 & \text{, if } t_i = 0 \end{cases} \text{, } i\in I. \qquad (18)$$

A value of T_i for which the infimum in (16) is attained can easily be evaluated for any given multipliers L that satisfy (18) as

$$\hat{T}_i(L) = \begin{cases} \dfrac{t_i}{1 + \sum_{a\in A} n_{ia} L_a} & \text{, if } t_i > 0 \\ 0 & \text{, if } t_i = 0 \end{cases} \text{, } i\in I. \qquad (19)$$

By inserting (19) into (16) and dropping the constant terms, we obtain a new formulation of the dual (15) which is no longer explicitly dependent on the primal variables T_i, but only on the dual variables L_a

$$\text{Min } g(L) = \sum_{i \in I^+} t_i \log \frac{t_i}{1 + \sum_{a \in A} n_{ia} L_a} + \sum_{a \in A} L_a v_a \qquad (20)$$

subject to (18).

Let T^* and L^* denote optimal solutions to the primal problem (8) and dual problem (20), respectively. (T^*, L^*) must satisfy the Kuhn-Tucker optimality conditions

$$1 + \sum_{a \in A} n_{ia} L_a^* \begin{cases} = \dfrac{t_i}{T_i^*} & , \text{ if } T_i^* > 0 \\ \\ \geq 0 & , \text{ if } T_i^* = 0 \end{cases} \quad , i \in I. \qquad (21)$$

which in this case are sufficient as well, since the problem is convex.

Proposition 4: The optimality conditions (21) are equivalent to

$$T_i^* \left(1 + \sum_{a \in A} n_{ia} L_a^*\right) = t_i \quad , i \in I. \qquad (22)$$

Proof: We first prove that (21) implies (22). For $i \in I^0$ (21) are complementarity conditions and (22) may be obtained by multiplying the corresponding terms. For $i \in I^+$ we note that only $T_i^* > 0$ can occur, since otherwise the objective function (8) becomes infinite, hence contradicting the optimality of T^*. Therefore only the first part of (21) applies to $i \in I^+$, which is, of course, equivalent to (22).

We now prove that (22) implies (21). For $T_i^* > 0$ (21) is obtained by dividing (22) by T_i^*. For $i \in I^o$, (21) is always satisfied, since L^* is dual feasible and therefore satisfies (18). ∎

3. A Solution Algorithm

Matrix estimation problems that occur in practice are typically of a very large size (hundreds of links and thousands of O-D pairs). Solution algorithms that require access to all data at the same time may be very difficult to implement in practice. In this section we discuss a simple algorithm, which is based on the idea of successive relaxation of all but one of the constraints (2). Since in this way the problem is broken down into smaller subproblems, each associated with one network link only, this kind of method is an attractive approach to large scale matrix estimation problems. The method is similar in spirit to the balancing methods used to solve the maximum entropy model (4), such as the algorithm proposed by Murchland [8]. But, in contrast to the maximum entropy model, Bregman's very general proof of convergence [1,6] may not be applied to our model, since the problem we consider is not contained in the class of problems considered by Bregman.

First, let us assume that $t_i > 0$ is satisfied for all O-D pairs $i \in I$ and consider the following algorithm:

Algorithm 1:

STEP 0: Set $L_a := 0$, $a \in A$.

STEP 1: For each $a \in A$ do:

a) Calculate

$$L_a^{min} = - \min_{\substack{i \in I \\ n_{ia} > 0}} \frac{1 + \sum_{b \neq a} n_{ib} L_b}{n_{ia}} \qquad (23)$$

b) Find $\bar{L}_a \in (L_a^{min}, \infty)$ that satisfies

$$\sum_{i \in I} \frac{n_{ia} t_i}{1 + \sum_{b \neq a} n_{ib} L_b + n_{ia} \bar{L}_a} = V_a. \qquad (24)$$

c) Set $d_a := \bar{L}_a - L_a$ and $L_a := \bar{L}_a$.

STEP 2: If $||d|| > \epsilon$ then go to STEP 1, else compute

$$T_i^* = \frac{t_i}{1 + \sum_{a \in A} n_{ia} L_a}, \quad i \in I. \qquad (25)$$

and STOP.

Proposition 5: For $t_i > 0$, $i \in I$, Algorithm 1 converges to the optimal solution of problem (8).

Proof: It suffices to show that the values of L_a, $a \in A$ converge to an optimal solution of the dual problem (20). Since the matrix (t_i) is positive, (18) contains only strict inequalities. The optimal solution of (20) is therefore an interior point of the feasible domain and unconstrained optimization techniques may be applied. Step 1 of Algorithm 1 corresponds to one iteration of the cyclic coordinate descent method applied to problem (20). The one-dimensional subproblem derived from (20) by fixing the values of all dual variables except for L_a is

$$\text{Min } g(L_a) = \sum_{i \in I} t_i \log \frac{t_i}{1 + \sum_{b \neq a} n_{ib} L_b + n_{ia} L_a} + \sum_{b \neq a} L_b V_b + L_a V_a. \qquad (26)$$

In step 1a) (23) assures that L stays dual feasible. Note

that $\frac{dg(L_a)}{dL_a}$ is a rational function with poles at $L_a = -\frac{1+\sum_{b\neq a} n_{ib}L_b}{n_{ia}}$ for all $a\epsilon A$ with $n_{ia}>0$. Restricting L_a to the interval (L_a^{min},∞) assures that in (24) $\frac{dg(L_a)}{dL_a}=0$ has exactly one solution, the one that satisfies (18). Since the cyclic coordinate descent method is well known to be globally convergent for unconstrained convex problems [7], the values of L_a, $a\epsilon A$ converge to an optimal solution of (20). The optimal values of the primal variables T can then be obtained by applying the optimality conditions (22), which is done in step 2 of the algorithm. ■

While Algorithm 1 is intended to be used only in the case $t_i>0$, $i\epsilon I$, it is still possible to use it even if some of the observed demands are zero. This is done by simply applying Algorithm 1 to the matrix (t_i') obtained from (t_i) by replacing all the zeroes by ϵ, a very small positive number.

It is possible to apply the cyclic coordinate descent method directly to the case in which I^o is not empty, which results in the following algorithm.

Algorithm 2:

STEP 0: Set $L_a := 0$, $a \in A$. Set $s_i := 1$, $i \in I$.

STEP 1: For each $a \in A$ do:

a) Calculate $d_a^{min} = - \min_{\substack{i \in I \\ n_{ia} > 0}} \frac{s_i}{n_{ia}}$ and let $\bar{I}$ be the set of O-D pairs for which this minimum is attained.

b) If $\bar{I} \subseteq I^o$ and $\sum_{i \in I^+} \frac{n_{ia} t_i}{s_i + n_{ia} d_a^{min}} \leq V_a$

then set $d_a = d_a^{min}$,

else find $d_a > d_a^{min}$ that satisfies

$$\sum_{i \in I^+} \frac{n_{ia} t_i}{s_i + n_{ia} d_a} = V_a .$$

c) Set $L_a := L_a + d_a$ and $s_i := s_i + n_{ia} d_a$, $i \in I$.

STEP 2: If $\|d\| > \epsilon$ then go to STEP 1, else compute

$$T_i^* = \frac{t_i}{s_i} \quad , \text{ for all } i \in I \text{ with } s_i > 0.$$

STEP 3: Calculate T_i^* for i with $s_i = 0$ by finding a feasible solution of the following linear system

$$\sum_{\substack{i \in I \\ s_i = 0}} n_{ia} T_i = V_a - \sum_{i \in I^+} n_{ia} T_i^* \quad , a \in A, \qquad (27)$$

$$T_i \geq 0 \quad , \text{ for all } i \in I^o \text{ with } s_i = 0.$$

and **STOP.**

Algorithm 2 is stated in a way that is somewhat closer to a computer code implementation than Algorithm 1. It is easy to see that we always have $s_i = 1 + \sum_{a \in A} n_{ia} L_a$, and that in the case in which $t_i > 0$, $i \in I$ the two algorithms are identical. As before, step 1 corresponds to one iteration of the cyclic coordinate descent method. Since in this case it is applied to a constrained problem, the convergence properties that hold in the unconstrained case are no longer valid. It is theoretically possible that Algorithm 2 terminates at an L that is not an optimal solution of problem (20). It is, however, possible to prove the following.

Proposition 6: If in step 3 of Algorithm 2 the feasible set of (27) is not empty, the algorithm stops at an optimal solution.

Proof: Let L^* denote the values of L when step 3 of the algorithm is commenced. It is evident that L^* is dual feasible. If T^* can be constructed, that is, if (27) has a solution, the resulting T^* is primal feasible. Since (T^*, L^*) satisfy the optimality conditions (22) by construction, they are optimal. ■

The feasibility of (27) can thus be used to determine whether the algorithm has properly converged to an optimal solution. Our experience shows that Algorithm 2 is very robust and that in all practical problems we experimented with it always converged. For certain classes of problems,

such as the matrix estimation problem with given marginal totals, (27) may be shown to be always feasible, in which case Algorithm 2 is globally convergent.

4. The Matrix Estimation Problem with Given Marginal Totals

A very important special case of the matrix estimation problem (1) arises when the marginal row and column totals of the O-D matrix are known. Let O_p denote the total production (row sum) of the origin zone p, $p \epsilon P$, and let D_q denote the total attraction of the destination zone q, $q \epsilon Q$. If again we make the assumption that the observed sample matrix (t_{pq}) consists of Poisson distributed integers, and that further the true values of the productions and attractions are known, we can again apply the maximum likelihood method to estimate the true O-D matrix (T_{pq}). In this case, (8) may be rewritten as

$$\text{Min} \sum_{p \epsilon P} \sum_{q \epsilon Q} (T_{pq} - t_{pq} log T_{pq}) \tag{28}$$

subject to

$$\sum_{q \epsilon Q} T_{pq} = O_p \quad , \; p \epsilon P, \tag{29}$$

$$\sum_{p \epsilon P} T_{pq} = D_q \quad , \; q \epsilon Q, \tag{30}$$

$$T_{pq} \geq 0 \quad , \; p \epsilon P, \; q \epsilon Q. \tag{31}$$

Without any loss of generality we can assume that $O_p > 0$, $p \epsilon P$, and $D_q > 0$, $q \epsilon Q$, since any origin p with $O_p = 0$, or destination q with $D_q = 0$ would simply result in a row or column of zeros in the matrix (T_{pq}). In order to assure feasibility of (29) and (30), the sums of the productions O_p must equal

the sums of the attractions D_q

$$\sum_{p\epsilon P} O_p = \sum_{q\epsilon Q} D_q = \sum_{p\epsilon P} \sum_{q\epsilon Q} T_{pq}. \qquad (32)$$

From (32) it follows that the first term in the objective function (28) is constant and may therefore be disregarded, yielding

$$\text{Max} \sum_{p\epsilon P} \sum_{q\epsilon Q} t_{pq} log T_{pq} \qquad (33)$$

subject to (29) - (31).

Let a_p, $p\epsilon P$, be the dual variables associated with production constraints (29) and b_q, $q\epsilon Q$, those associated with the attraction constraints (30). Using a similar development as the one that led to (20), we can formulate the Lagrangean dual of (33) as

$$\text{Min} \sum_{p\epsilon P} \sum_{q\epsilon Q} t_{pq} log \frac{t_{pq}}{a_p + b_q} + \sum_{p\epsilon P} a_p O_p + \sum_{q\epsilon Q} b_q D_q \qquad (34)$$

subject to

$$a_p + b_q \geq 0 \qquad , p\epsilon P,\ q\epsilon Q. \qquad (35)$$

Let T^* be an optimal solution of the primal problem (33) and let (a^*, b^*) be an optimal solution of the dual (34). The Kuhn-Tucker optimality condition for problem (33) is

$$T^*_{pq}(a^*_p + b^*_q) = t_{pq} \qquad , p\epsilon P,\ q\epsilon Q, \qquad (36)$$

which for $t_{pq}>0$ corresponds to the following simple model

$$T^*_{pq} = \frac{t_{pq}}{a^*_p + b^*_q}. \qquad (37)$$

Note that this model, resulting from the maximum likelihood approach in the case of Poisson distributed observed sample matrix $\{t_i\}$, is different from the maximum entropy model, also known as biproportional balancing, Fratar or Furness method [3,6], for which we have

$$T^*_{pq} = e^{a^*_p + b^*_q} t_{pq}, \quad p\epsilon P, \ q\epsilon Q. \qquad (38)$$

Algorithm 2 of the previous section can be adapted to the matrix estimation problem (33).

Algorithm 3:

STEP 0: Set $a_p := 1/2$, $p\epsilon P$, and $b_q := 1/2$, $q\epsilon Q$.

STEP 1: For each $p\epsilon P$ do:

Find $\bar{a}_p \geq -\min_{q\epsilon Q} b_q$ that minimizes $\left| \sum_{q\epsilon Q} \frac{t_{pq}}{\bar{a}_p + b_q} - O_p \right|$.

Set $d_p := a_p - \bar{a}_p$ and $a_p := \bar{a}_p$.

STEP 2: For each $q\epsilon Q$ do:

Find $\bar{b}_q \geq -\min_{p\epsilon P} a_p$ that minimizes $\left| \sum_{p\epsilon P} \frac{t_{pq}}{a_p + \bar{b}_q} - D_q \right|$.

Set $d_q := b_q - \bar{b}_q$ and $b_q := \bar{b}_q$.

STEP 3: If $||d|| > \epsilon$ then go to STEP 1, else compute

$$T^*_{pq} = \frac{t_{pq}}{a_p + b_q}, \quad (p,q)\epsilon I^+.$$

STEP 4: Compute

$$T^*_{pq} = \frac{\bar{O}_p \bar{D}_q}{\bar{T}} \quad , (p,q) \epsilon I^O, \tag{39}$$

where

$$\bar{O}_p = O_p - \sum_{\substack{q \epsilon Q \\ t_{pq} > 0}} T^*_{pq} \quad , p \epsilon P, \tag{40}$$

$$\bar{D}_q = D_q - \sum_{\substack{p \epsilon P \\ t_{pq} > 0}} T^*_{pq} \quad , q \epsilon Q, \tag{41}$$

$$\bar{T} = \sum_{p \epsilon P} \bar{O}_p = \sum_{q \epsilon Q} \bar{D}_q. \tag{42}$$

and **STOP.**

The one-dimensional minimization problems in step 1 and 2 are equivalent to finding the zero of the corresponding function, if it exists. If the function has no zero-crossing in the considered interval, the minimum is always attained at the lower bound.

Proposition 7: Algorithm 3 converges to an optimal solution of (33).

Proof: Since problem (33) is a special case of the matrix estimation problem (8) and the algorithm is equivalent to Algorithm 2 applied to (33), we can apply the result of proposition 6. Hence, we have only to show that the linear system corresponding to (27) is always feasible. In the context of given marginal totals (27) translates to

$$\sum_{q\in\bar{Q}} T_{pq} = \bar{O}_p \qquad , p\in\bar{P}, \tag{43}$$

$$\sum_{p\in\bar{P}} T_{pq} = \bar{D}_q \qquad , q\in\bar{Q}, \tag{44}$$

$$T_{pq} \geq 0 \qquad , p\in\bar{P},\ q\in\bar{Q}. \tag{45}$$

The residual productions $\bar{O}_p$ and attractions $\bar{D}_q$ are defined in (40) and (41). $\bar{P}$ is the set of origins with $\bar{O}_p>0$ and $\bar{Q}$ is the set of destinations with $\bar{D}_q>0$. The constraints (43) - (45) are of the same structure as the original constraints (29) - (31), but are only applied to a submatrix. The system (43) - (45) is therefore feasible if and only if

$$\sum_{p\in\bar{P}} \bar{O}_p = \sum_{q\in\bar{Q}} \bar{D}_q$$

As can be verified easily using (40) and (41), this condition is always satisfied. Hence the resulting T^* is primal feasible. Since L is dual feasible and the optimality conditions (36) are satisfied by construction, we conclude that T^* is an optimal solution of (33). ■

Note that in general, the linear system (43) - (45) may have more than one feasible solution. The solution calculated in step 4 of the algorithm is the one that maximizes the entropy function

$$\text{Max} \sum_{p\in\bar{P}} \sum_{q\in\bar{Q}} T_{pq} \log T_{pq}.$$

Of course, any other feasible solution of (43) - (45) will yield an optimal solution of (33) as well.

Having established the convergence of the algorithm, we now discuss briefly how the one dimensional subproblems that arise in step 1 and 2 can be solved efficiently. As we have already seen, the main problem consists of finding the zero-crossings of the corresponding functions. This can be done efficiently by applying Newton's method (successive linear approximation), which results in the following recursion formulas

$$\text{in step 1: } \bar{a}_p := \bar{a}_p + \frac{\sum_{q \in Q} \frac{t_{pq}}{\bar{a}_p + b_q} - O_p}{\sum_{q \in Q} \frac{t_{pq}}{(\bar{a}_p + b_q)^2}} \tag{46}$$

and

$$\text{in step 2: } \bar{b}_q := \bar{b}_q + \frac{\sum_{p \in P} \frac{t_{pq}}{a_p + \bar{b}_q} - D_q}{\sum_{p \in P} \frac{t_{pq}}{(a_p + \bar{b}_q)^2}} \tag{47}$$

Convergence of the above recursions is assured if the following initial values are used

$$\bar{a}_p := \frac{\sum_{q \in Q^{min}} t_{pq}}{O_p} - b^{min} \quad \text{and} \quad \bar{b}_q := \frac{\sum_{p \in P^{min}} t_{pq}}{D_q} - a^{min} \tag{48}$$

where P^{min} and Q^{min} are the sets of origins and destinations for which the minimum of a_p and b_q is attained. However, since after some iterations the values of the dual variables

change very slowly, it is usually faster to start with the current value of the dual variables and to use the above initial value only if the recursion leads to values outside the feasible region.

The following small example with only 4 origins and 4 destinations is used to illustrate Algorithm 3. Productions O_p, attractions D_q and the observed sample matrix $\{t_{pq}\}$ are given in Table 1.

t_{pq}	q=1	2	3	4	O_p
p=1	10.000	.000	15.000	5.000	24.000
2	.000	11.000	.000	.000	18.000
3	6.000	9.000	5.000	.000	19.000
4	.000	.000	8.000	4.000	11.000
D_q	19.000	17.000	23.000	13.000	72.000

Table 1: Small example problem

Table 2 shows the values of the dual variables at each iteration, as well as the value of the dual objective function and the stopping criterion $\|d\|$. Underlined values of a_p and b_q correspond to active lower bounds.

Iter.	a1	a2	a3	a4	b1	b2	b3	b4	Obj.Fct.	$\|d\|$
0	.500	.500	.500	.500	.500	.500	.500	.500	230.8241	-
1	.750	.111	.553	.591	.177	.905	.554	.022	224.3636	2.042
2	.949	-.022	.548	.778	.091	1.004	.409	.022	222.4997	.855
3	1.035	-.022	.612	.844	.022	.989	.336	.022	222.2466	.373
4	1.090	-.022	.673	.881	.022	.975	.284	.022	222.1411	.219
5	1.110	-.022	.688	.909	.022	.972	.263	.022	222.1265	.087
6	1.119	-.022	.694	.921	.022	.971	.255	.022	222.1240	.036
7	1.122	-.022	.697	.926	.022	.970	.251	.022	222.1236	.015
8	1.124	-.022	.698	.928	.022	.970	.249	.022	222.1235	.006
9	1.124	-.022	.698	.928	.022	.970	.249	.022	222.1235	.003
10	1.125	-.022	.698	.929	.022	.970	.248	.022	222.1235	.001

Table 2: Dual varibles, dual objective function and stopping criterion

Finally, in Table 3 the optimal solution calculated in steps 3 and 4 is displayed.

T_{pq}^*	q=1	2	3	4	O_p
p=1	8.718	.000	10.924	4.359	24.000
2	1.958*	11.606	.000	4.436*	18.000
3	8.326	5.394	5.281	.000	19.000
4	.000	.000	6.795	4.205	11.000
D_q	19.000	17.000	23.000	13.000	

Table 3: Optimal Solution T_{pq}^* ($*= (pq) \in \bar{P}x\bar{Q}$)

It is interesting to note that the Furness method could not be applied to the above example, since the resulting biproportional model (38) is infeasible, due to the many O-D pairs with $t_{pq}=0$.

5. Statistical Inference

Little work has been reported so far on statistical methods that may be used to test the consistency of the data used in O-D matrix estimations. As we have seen in Section 1, in the existing matrix estimation models $\{t_i\}$ is defined to be a *constant* reference matrix of unspecified statistical provenance. They are aimed at finding the matrix $\{T_i\}$ that is the most similar to $\{t_i\}$, among all matrices satisfying (2) and (3). The statistical aspect of these models is therefore limited to measures of goodness of fit, such as those proposed by Smith and Hutchinson [10]; however, it is not possible to qualify statistically the estimates that are obtained from these model.

In the preceding sections we have always *assumed* that the elements of the sample matrix $\{t_i\}$ are Poisson distributed with means $\{T_i\}$ which satisfy (2). This implies that we know a) the *true* link volumes V_a and b) the *true* link proportions n_{ia}. In most application we do have reasonable estimates of these quantities, but we can by no means guarantee that they are the *true* values. It would therefore be extremely helpful to have statistical tests that allow us to test the null-hypothesis H0: *T* satisfies (2), versus the alternative H1: *T* does not satisfy (2).

A simple statistic that allows us to test the above hypothesis is based on the quantity

$$v_a = \sum_{i \epsilon I} n_{ia} t_i \qquad (49)$$

and is applicable whenever the link proportions n_{ia} are limited to the values 0 and 1. In this case it is easy to verify that v_a is Poisson distributed with mean V_a. H0 is therefore rejected if v_a lies outside the confidence of the corresponding Poisson distribution.

Note that this test is for one link of the network only. Since several tests of this type, for different links, are in general not mutually independent, care must be taken when interpreting the results of several tests. In the case of the matrix estimation problem with given marginal totals, tests on production constraints (attraction constraints) are mutually independent. In this case the quantities

$$o_p = \sum_{q \epsilon Q} t_{pq} \qquad , p \epsilon P, \qquad (50)$$

are independently Poisson distributed with means O_p, and

$$d_q = \sum_{p \epsilon P} t_{pq} \qquad , q \epsilon Q, \qquad (51)$$

are independently Poisson distributed with means D_q. Note that the distributions of o_p and d_q are not independent, since they both include the trips t_{pq}.

A more comprehensive test is based on the likelihood ratio λ, that is the maximum value of the likelihood (6) if constraints (2) are imposed, divided by the maximum value of the likelihood if no constraints are imposed. It is well

known (see e.g. Wilks [15]) that $-2log\lambda$ is asymptotically chi-square distributed with as many degrees of freedom as there are non-redundant constraints in (2). Since for the matrix estimation problem (8) the optimum of the unconstrained problem is attained at $\{t_i\}$, we have

$$Y = -2log\lambda = 2 \sum_{i \in I} (T_i^* - t_i (1 + log \frac{T_i^*}{t_i})). \qquad (52)$$

Asymptotically, Y is chi-square distributed with $|A|$ degrees of freedom, supposing that (2) contains no redundant constraints.

For the matrix estimation problem with given marginal totals we have analogously that Y is asymptotically chi-square with $|P|+|Q|-1$ degrees of freedom, since (32) implies that there is always one redundant constraint in (29) and (30). The statistic Y is very similar to a measure of goodness of fit proposed by Wilson [15] and studied by Smith and Hutchinson [10] in the context of trip distribution models. Smith and Hutchinson used the measure

$$LLD = \sum_{p \in P} \sum_{q \in Q} t_{pq} log \frac{t_{pq}}{T_{pq}^*} \qquad (53)$$

and named it "log likelihood difference". The two statistics Y and LLD are, however, not equivalent unless $\sum_{p \in P} \sum_{q \in Q} t_{pq} = \sum_{p \in P} \sum_{q \in Q} T_{pq}^*$ is satisfied.

Due to the complexity of the model, the exact

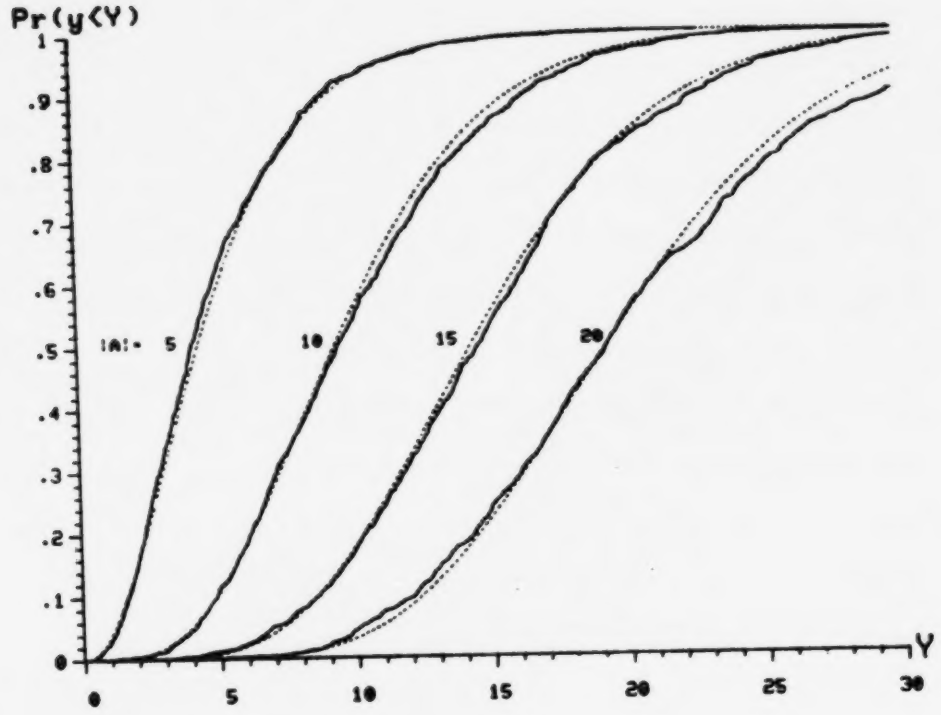

Figure 1: Distribution of the statistic Y for the matrix estimation problem with given link volumes. The empirical distributions (solid lines) obtained after 1000 trials are compared to the asymptotic chi-square distributions (dotted lines).

distribution of the statistic Y can not be calculated analytically. Nevertheless, since the problems usually considered are quite large and involve many individual trips, we expect the actual distribution of Y to be approximated well by the asymptotic chi-square distribution. In order to verify empirically the similarity of the actual and asymptotic distributions, the Monte Carlo simulation technique was used to generate values of y_i that follow the exact distribution of Y for some sample problems. The examples

chosen were linear networks (link chains) with 5, 10, 15 and 20 links, corresponding to 15, 55, 120 and 210 O-D pairs. A true O-D matrix T was assumed and the link volumes were calculated accordingly. Next, for each trial a sample matrix $\{t_i\}$ was generated using Poisson distributed random numbers. Problem (8) was solved by using Algorithm 2 and the value of Y was computed. Figure 1 shows the comparison of the empirical distributions obtained after 1000 trials and the asymptotic chi-square distributions for each example problem. The Kolmogorov-Smirnov test [16] was applied to the distributions of Figure 1 and no significant difference between the actual and chi-square distribution was detectable at a 5% error level.

The same Monte Carlo simulation technique was also used for the case of the matrix estimation problem with given marginal totals. Five different "true" O-D matrices with the dimensions $|P|=|Q|=5$, 10, 15, 20 and 25 were generated by using exponentially distributed random numbers with a mean of 5 trips. Figure 2 shows the empirical distributions obtained after 1000 trials, compared to the asymptotic chi-square distribution. Again, the Kolmogorov-Smirnov test did not indicate any significant difference in all five cases.

All the experiments carried out seem to indicate that the statistic Y is approximated very well by a chi-square distribution whenever the link volumes V_a (or the productions O_p and attractions D_q) are not too small (say larger

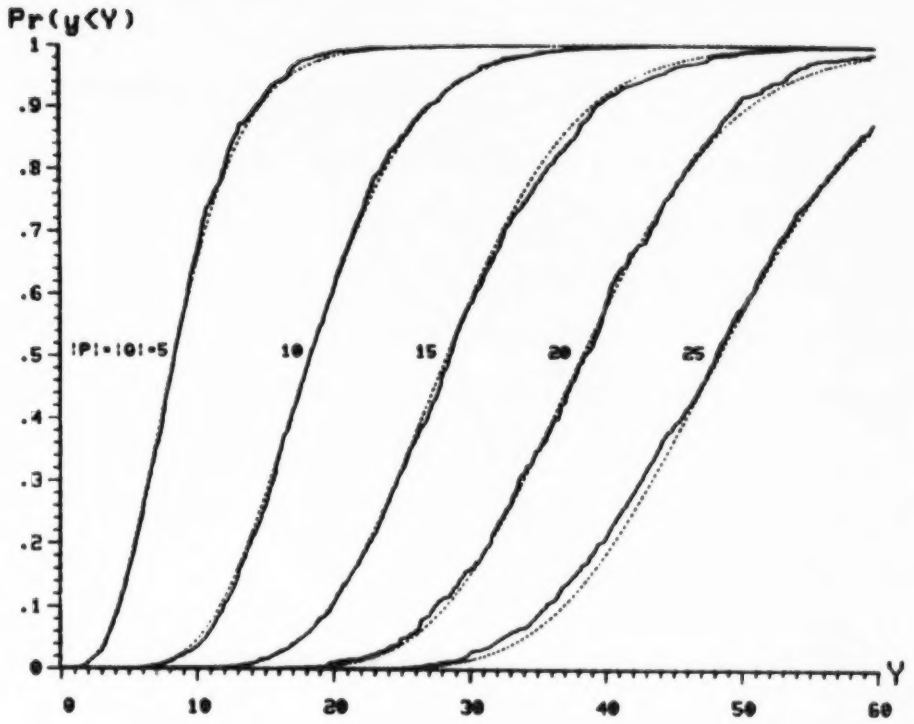

Figure 2: Distribution of the statistic Y for the matrix estimation problem with given marginal totals. The empirical distributions (solid lines) obtained after 1000 trials are compared to the asymptotic chi-square distributions (dotted lines).

than 10). Figure 3 shows how the distribution of Y behaves for smaller values of V_a or D_p and D_q. Matrix estimation problems with |P|=|Q|=10 were generated with mean productions and attractions of 2, 5, 10 and 25 trips. As we can see from Figure 3, the chi-square approximation, indicated with dotted lines, no longer holds for values smaller than 10. Since realistic values of link volumes or productions/attractions are usually well above this critical value (typically in the hundreds or thousands) the chi-

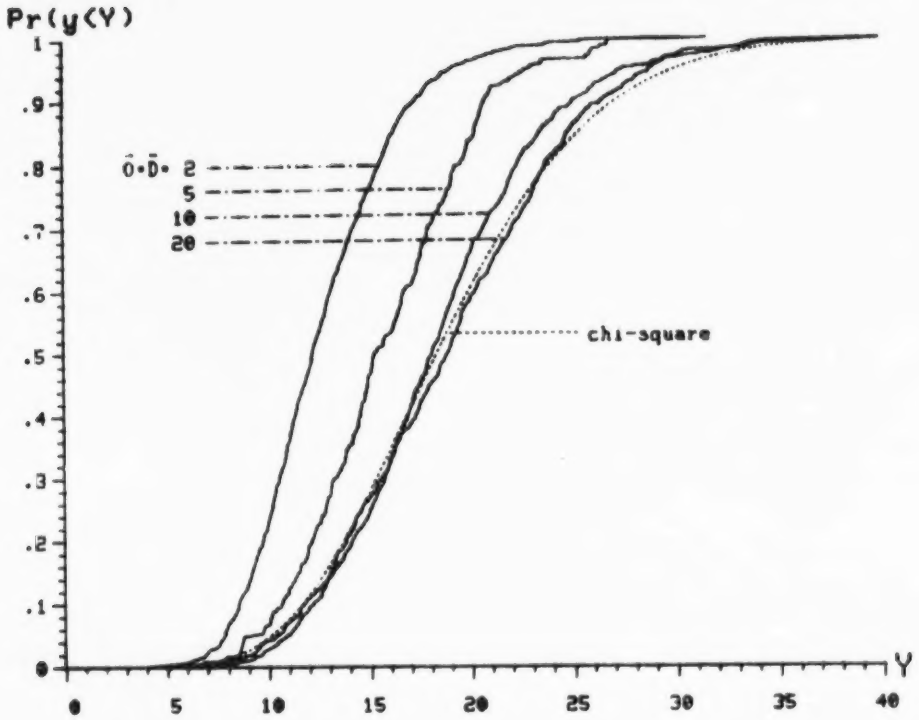

Figure 3: Distribution of Y for small values of productions and attractions. Chi-square approximation is only valid for values larger than 10.

square approximation can be used for virtually all problems that occur in practice.

6. Model Extensions

So far we have always assumed that the true values of the link volumes (or productions and attractions) are known. This is a reasonable assumption in all cases in which the volumes are counted on a regular basis and therefore their mean values can be estimated reliably. It is however possible, that the only information available on the link volumes are the counts obtained from a one-time survey. To assume that these counts represent the true link volumes would not be reasonable. In this case, a more realistic model results if we interpret the observed volumes, denoted v_a, to be random samples of a Poisson distribution with the (unknown) mean V_a. If the observed volumes v_a were obtained from mutually independent surveys and independently from the sample O-D matrix (t_i), the maximum likelihood method yields the following model

$$\text{Min} \sum_{i \in I} (T_i - t_i \log T_i) + \sum_{a \in A} (V_a - v_a \log V_a) \qquad (54)$$

subject to (2) and (3). Clearly, (54) is still a convex programming problem. The Kuhn-Tucker conditions for an optimal solution of (54) are (22) and

$$v_a = V_a^*(1-L_a^*) \quad , \; a \in A. \qquad (55)$$

If $v_a > 0$ (55) can be used to substitute V_a by L_a or vice versa. Algorithms 1 and 2 may be modified to solve (54). For Algorithm 1 this is simply done by replacing V_a by

$v_a/(1-\bar{L}_a)$ in (24) and, having $\bar{L}_a$ restricted to the interval $(L_a^{min},1)$ in step 1 in order to ensure non-negativity of V_a^*.

This modified version of the O-D matrix estimation problem is especially useful if the available observed link volumes v_a, $a\epsilon A$, do not satisfy the conservation of flow equations and would therefore, if used in model (8), result in an inconsistent set of constraints (2). The extended model (54) is always feasible and the resulting set of estimated link volumes V_a, $a\epsilon A$, will always satisfy the conservation of flow equations.

Since (54) is also a maximum likelihood model, we can, as was done in the preceding section, calculate the statistic Y as

$$Y = -2 \sum_{i\epsilon I} (T_i^* - t_i (1+log\frac{T_i^*}{t_i})) -2 \sum_{a\epsilon A} (V_a^* - v_a (1+log\frac{V_a^*}{v_a})) \quad (56)$$

which is asymptotically chi-square distributed with |A| degrees of freedom. Monte Carlo simulations were carried out for the same example networks as were used in the preceding section. The resulting empirical distributions are shown in Figure 4. Again, no significant difference between the empirical and the asymptotic chi-square distributions are apparent.

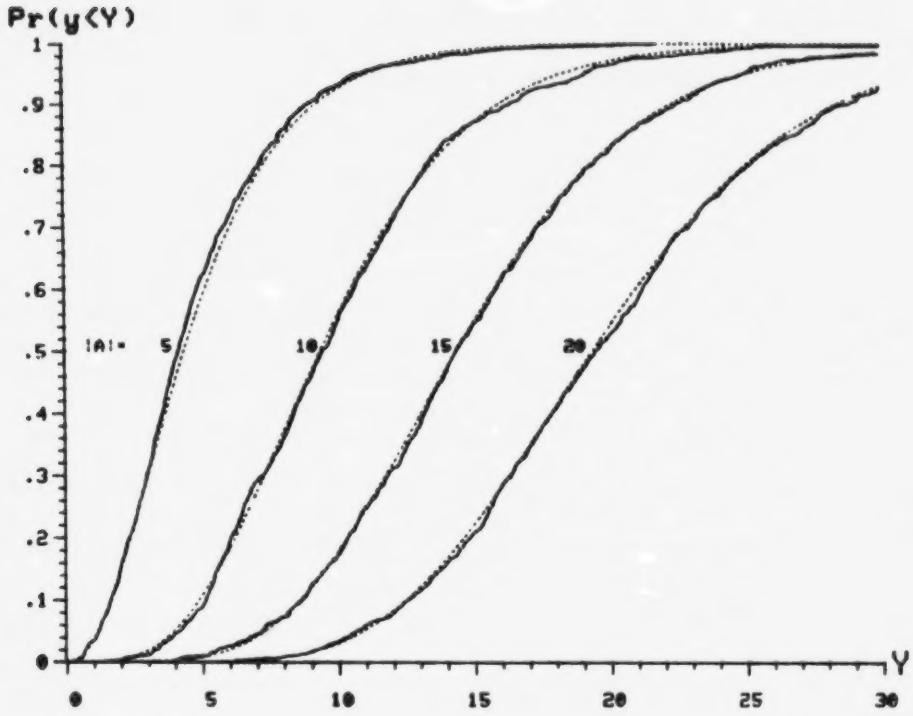

Figure 4: Distribution of the statistic Y for the matrix estimation problem with observed link volumes. The empirical distributions (solid lines) obtained after 1000 trials are compared to the asymptotic chi-square distributions (dotted lines).

The same extension applied to the special case of the matrix estimation problem with constraints on the marginal totals yields a model of the following form

$$\text{Min} \sum_{\substack{p\in P\\ q\in Q}} (T_{pq} - t_{pq} log T_{pq}) + \sum_{p\in P} (O_p - o_p log O_p) + \sum_{q\in Q} (D_q - d_q log D_q) \quad (57)$$

subject to (29) - (31). The values o_p and d_q are the observed productions and attractions, obtained by mutually independent counts and independently from the sample matrix

$\{t_{pq}\}$. If t_{pq}, o_p and d_q are all positive, the optimality conditions of (57) can be stated as

$$\frac{t_{pq}}{T^*_{pq}} + \frac{o_p}{O^*_p} + \frac{d_q}{D^*_q} = 3 \qquad , p\epsilon P,\ q\epsilon Q. \tag{58}$$

An algorithm similar to Algorithm 3 may be used to solve this modified matrix estimation problem with observed marginal totals. The following algorithm assumes that all observed productions and attractions are strictly positive; a slightly more complicated algorithm would result if this assumption were not satisfied.

Algorithm 4:

STEP 0: Set $a_p := 1/2$, $p\epsilon P$, and $b_q := 1/2$, $q\epsilon Q$.

Set $c = \left(\sum_{p\epsilon P} o_p - \sum_{q\epsilon Q} d_q\right) / \left(\sum_{p\epsilon P} o_p + \sum_{q\epsilon Q} d_q\right)$.

STEP 1: For each $p\epsilon P$ do:

Find $\bar{a}_p$ that satisfies $3/2+c \geq \bar{a}_p \geq -\min_{q\epsilon Q} b_q$ and minimizes

$$\left| \sum_{q\epsilon Q} \frac{t_{pq}}{\bar{a}_p + b_q} - \frac{o_p}{3/2 - \bar{a}_p + c} \right|.$$

Set $d_p := a_p - \bar{a}_p$ and $a_p := \bar{a}_p$.

STEP 2: For each $q\epsilon Q$ do:

Find $\bar{b}_q$ that satisfies $3/2-c \geq \bar{b}_q \geq -\min_{p\epsilon P} a_p$ and minimizes

$$\left| \sum_{p\epsilon P} \frac{t_{pq}}{a_p + \bar{b}_q} - \frac{d_q}{3/2 - \bar{b}_q - c} \right|.$$

Set $d_q := b_q - \bar{b}_q$ and $b_q := \bar{b}_q$.

STEP 3: Find $\bar{c}$ in the interval $-3/2+\max_{p\in P} a_p \leq \bar{c} \leq 3/2-\max_{q\in Q} b_q$ that satisfies

$$\sum_{p\in P} \frac{o_p}{1-a_p+\bar{c}} - \sum_{q\in Q} \frac{d_q}{1-b_q-\bar{c}} = 0$$

Set $d_0 := c - \bar{c}$ and $c := \bar{c}$.

STEP 4: If $\|d\| > \epsilon$ then go to STEP 1, else compute

$$T^*_{pq} = \frac{t_{pq}}{a_p+b_q} \quad , (p,q)\in I^+.$$

$$O^*_p = \frac{o_p}{3/2-a_p+c} \quad , p\in P,$$

$$D^*_q = \frac{d_q}{3/2-b_q-c} \quad , q\in Q,$$

STEP 5: Compute T^*_{pq}, $(p,q)\in I^o$, using (39) - (42) and STOP.

Algorithm 4 also corresponds to the cyclic coordinate descent method applied to the dual of (57). In order to ensure that the subproblem that has to be solved for T^*_{pq}, $(p,q)\in I^o$, is feasible, the redundant constraint (32) has been included explicitly into the model formulation and c is the corresponding dual variable. Note that step 3 is only necessary if I^o is not empty. We can therefore state

Proposition 8: Algorithm 4 converges to an optimal solution of problem (57).

The proof of proposition 8 is very similar to the proof

of proposition 7 and is therefore left to the reader.

For problem (57) the statistic Y is defined as

$$Y = 2\sum_{\substack{p\in P\\ q\in Q}} (3T^*_{pq} - t_{pq}(1+log\frac{T^*_{pq}}{t_{pq}})) - 2\sum_{p\in P} o_p(1+log\frac{O^*_p}{o_p}) - 2\sum_{q\in Q} d_q(1+log\frac{D^*_q}{d_q}) \quad (59)$$

and is asymptotically chi-square distributed with $|P|+|Q|$ degrees of freedom, since in this case (29) and (30) no longer contain a redundant constraint. Figure 5 shows the results of the Monte Carlo simulations for the same examples as were used in Figure 2. The empirical and chi-square distributions correspond very well. Note also that the shift of one degree of freedom is confirmed by the empirical distributions.

There is a practical problem associated with the extended models mentioned above, since the derivation of (54) is based on the assumption that the observed volumes v_a are obtained from mutually independent counts. In practice, this means that the observation (counting) of the volume on each link has to occur at different times, since otherwise the same individual trip could be counted on more than one link; this would contradict the assumption of the mutual independence of the observations. While model (54) may still be applied using link volumes that are not independent, that is, volumes resulting from counts in overlapping time periods, (54) can no longer be interpreted as a maximum likelihood model, but has to be classified as an

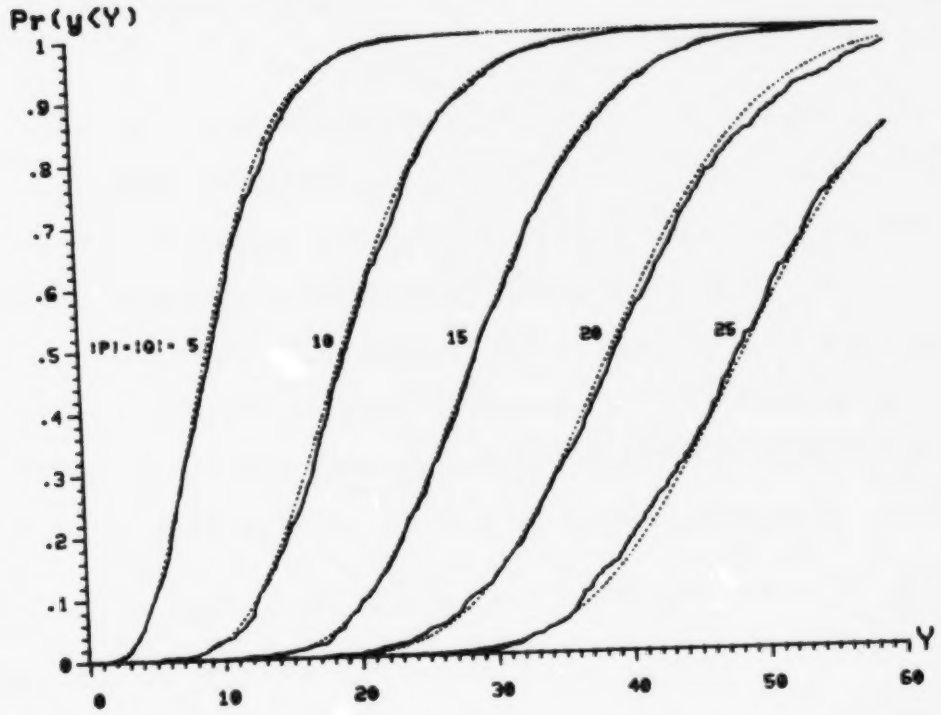

Figure 5: Distribution of the statistic Y for the matrix estimation problem with observed marginal totals. The empirical distributions (solid lines) obtained after 1000 trials are compared to the asymptotic chi-square distributions (dotted lines).

approximation thereof.

In the case of the matrix estimation problem with observed marginal totals (57), independence of the observations is assured as long as the productions were obtained independently from the attractions. In practice, this will be satisfied in many cases, e.g. if the productions are proportional to the working population living within a zone and the attractions are proportional to the number of jobs available in a zone, the data is usually obtained by two

completely independent surveys.

The same problem of the independence of the observed link volumes v_a arises also in a different context. Van Zuylen and Willumsen [14] and later Van Zuylen and Branston [13] study a maximum likelihood model to estimate a set of consistent link volumes. They also assume the observed link volumes to be Poisson-distributed and their model too, needs the assumption of mutual independence of the observed link volumes, even though this is not stated explicitly.

7. Conclusions

We have proposed a new model for estimating an O-D matrix when link volumes and a Poisson-distributed sample matrix is known. It was shown that applying the maximum likelyhood model results in a convex programming problem and an algorithm was developed to solve this model. The special case of the matrix estimation problem, when productions and attractions of the zones are known instead of link volumes, was considered separately.

Since the model we propose is based on the maximum likelihood method, this approach distinguishes itself from other matrix estimation approaches in that it is possible to assess the appropriateness of this model for a given set of data, by applying the likelyhood ratio test. We showed empirically using Monte Carlo simulations that, if the link volumes are reasonably large (>10), the resulting statistic Y is approximated very well by a chi-square distribution. The statistic Y may therefore be used to test the coherence between sample matrix, link volumes and link probabilities.

It is important to note that, in order to apply the maximum likelihood method, it is necessary to give a precise interpretation of the reference matrix $\{t_i\}$. In our model we assume the elements of $\{t_i\}$ to be independently Poisson-distributed. While other assumptions on the provenance of $\{t_i\}$ could be postulated and would yield different maximum

likelihood models, the case of Poisson-distributed observed O-D matrices is the most important one in practice, since it corresponds to most of the situations in which $\{t_i\}$ was obtained by *counting* trips.

A practical advantage of our model concerns the feasibility with respect to the values of $\{t_i\}$. In contrast to the maximum entropy model in which zero-elements in $\{t_i\}$ may render the model infeasible, our model is always feasible as long as the link volumes are consistent. In the extension of the model in which the link volumes are also Poisson-distributed observations, the feasibility of the model is assured even if the observed link volumes are not consistent.

Our experiments have shown that for all practical purposes, the statistic Y can be regarded as being chi-square distributed. So far, this is supported theoretically only by asymptotical results, that is, for the case in which a large number of observations of the sample matrix $\{t_i\}$ have been made. But in our model we suppose that we have only *one* observation of the sample matrix $\{t_i\}$. It is therefore surprising how well actual and chi-square distributions correspond. Further research is necessary to investigate the exact distribution of Y and the conditions necessary for applying the chi-square approximation.

References

1. Bregman L., "The Relaxation Method of Finding the Common Point of Convex Sets and its Application to the Solution of Problems in Convex Programming", U.S.S.R. Computational Math. Mathematical Phys. 7, pp 200-217, 1967.

2. Deming W.E. and Stephan F.F., "On a Least Square Adjustment of a Sampled Frequency Table when Expected Marginal Totals are Known", Annal of Mathematical Statistics, Vol 11, pp 427-444, 1940.

3. Furness K.P., "Time Function Interaction", Traffic Engineering and Control, Vol 7, No 7, pp 19-36, 1970.

4. Jörnsten K.O. and Nguyen S., "On the Estimation of a Trip Matrix from Network Data", Centre de Recherche sur les Transports, Universitè de Montrèal, Publication 153, 1979.

5. Kirby H.R., "Theoretical Requirements for Calibrating Gravity Models", Transportation Research, Vol 8, pp 97-104, 1974.

6. Lamond B. and Stewart N.F., "Bregman's Balancing Method", Transportation Research B, Vol 15B, pp 239-248, 1981.

7. Luenberger D.G., Introduction to Linear and Nonlinear Programming, Addison-Wesley, 1973.

8. Murchland J., "The Multi-proportional Problem", University College London, research note JDM 263, 1977.

9. Nguyen S., "Estimating Origin-Destination Matrices from Observed Flows", Proceedings of the First Course on Transportation Planning Models of the International School of Transportation Planning, Amalfi, Italy, October 1982.

10. Smith D.P. and Hutchison B.G., "Goodness of Fit Statistic for Trip Distribution Models", Transportation Research A, Vol 15A, pp 295-303, 1981.

11. Van Vliet D. and Willumsen L.G., "Validation of the ME2 Model for Estimating Trip Matrices from Traffic Counts", Proceedings of The Eighth International Symposium on Transportation and Traffic Theory, June 1981.

12. Van Zuylen H.J., "Some Improvements in the Estimation of an O-D Matrix from Traffic Counts", Proceedings of The Eighth International Symposium on Transportation and Traffic Theory, June 1981.

13. Van Zuylen H.J. and Branston D.M., "Consistent Link Flow Estimation from Counts", Transportation Research B, Vol 16B, pp 473-476, 1982.

14. Van Zuylen H.J. and Willumsen L.G, "The Most Likely Trip Matrix Estimated from Traffic Counts", Transportation Research B, Vol 14B, pp 281-293, 1980.

15. Wilks S.S., Mathematical Statistics, Wiley & Sons, New York, 1962.

16. Wilson S.R., "Statistical Notes on the Evaluation of Calibrated Gravity Models", Transportation Research, Vol 10, pp 343-345, 1976.

Acknowledgements

This paper was stimulated in part by a seminar presentation on O-D matrix estimation by Dirk Van Vliet at the Centre de recherche sur les transports. The autor benefitted from discussions with Sang Nguyen on the contents of this paper. Michael Florian and Linda James-Lefebvre assisted the author with stylistic and expository aspects of the final presentation.